STI Conference 2016 · València Peripheries, frontiers and beyond
València, 14 · 16 September 2016

Proceedings of the 21st International Conference on Science and Technology Indicators | València (Spain) | September 14-16, 2016
DOI: http://dx.doi.org/10.4995/STI2016.2016.xxxx

Capturing knowledge integration through collaborations: measures of the diversity and coherence in multiple proximity dimensions

By Frédérique Lang*, Ismael Ràfols[1]*, Michael Hopkins*
*f.lang@sussex.ac.uk
SPRU (Science and Policy Research Unit), University of Sussex, Brighton, UK
[1] Ingenio (CSIC-UPV), Universitat Politècnica de València, València, Spain

1. INTRODUCTION

This paper proposes a novel approach to determine changes that result from collaborations aimed at supporting knowledge integration. The approach combines and applies indicators of proximity, diversity, and coherence. It has potential applications in the study and evaluation of research collaborations.

The science studies literature has been exploring the topic of knowledge integration and interdisciplinarity for decades, both in scientometrics (Bordons, 2004; Zitt, 2005; Rafols, 2014), research management (Grant & Baden-Fuller, 1995; Nooteboom, 2000) and geography (Maskell & Malmberg, 1999). Contributions have extensively discussed how different bodies of knowledge are brought together within publication, individuals, university department or systems (Porter et al. 2007; Rafols et al. 2012 ; Zhou e al. 2012). The paper builds on the line of research that seeks to develop measures of knowledge integration, namely diversity and coherence (Rafols, 2014). Until now these measures of knowledge integration have mainly been used to look at disciplinary or cognitive differences.

The changes of successful exchange and integration of knowledge through collaboration is not only influenced by the disciplinary or cognitive diversity of the participants, as previously studied in the scientometric literature (Rafols et al., 2012; Rafols, 2014) but also other dimensions linked to the social, cultural background of those involved. A powerful framework has emerged from economic geography - the proximity framework (Boschma, 2005), that identifies five features that may be important for collaborative learning and hence collaborative knowledge integration which are: cognitive, social, geographical, institutional, and organisational proximities.

This paper therefore proposes to use diversity and coherence measures to not only look at diversity from a cognitive standpoint, but also apply it to the other proximities proposed in the Boschma framework. These indicators will capture the relationship occurring between individuals taking part in the research and the categories (proximity dimensions) that they are associated to.

This paper reviews and integrates concepts from economic geography with the scientometric literature on interdisciplinarity to form a conceptual framework that the paper applies to an illustrative case study. In order to apply the framework, the paper develops indicators for diversity and coherence that can be applied to each of Boschma's five proximities. The paper proposes to use diversity and coherence measures with the five dimensions, and it also uses the

elements behind the diversity and coherence measures to create visualisations that has the potential to represent some complexity hidden behind the metrics. The illustrative case study looks at collaborations between individuals within a biomedical research project on Podoconiosis. The method aims to build not only indicators to look at diversity in collaboration, but also new ways of mapping relationships using the proximity framework.

2. ANALYTICAL FRAMEWORK

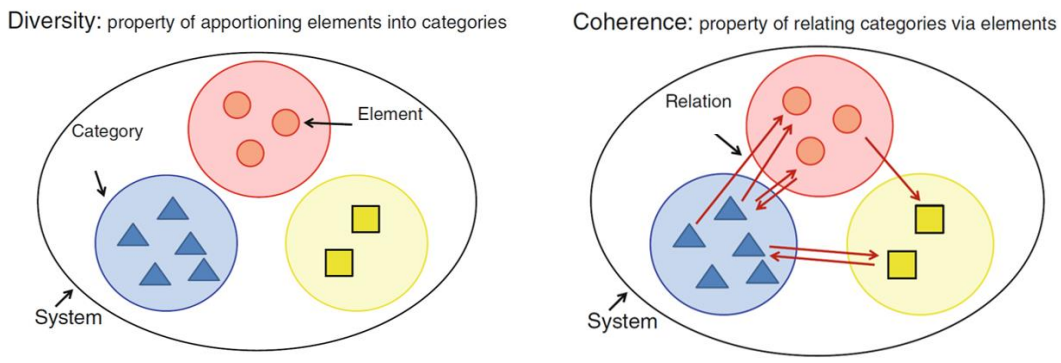
This section explores the framework developed by Rafols (Rafols & Meyer, 2010; Rafols, Porter, & Leydesdorff, 2010; Rafols, 2014) to study diversity, and shows how this can be associated to the theoretical framework developed by Boschma (2005) on the 5 proximities dimensions proposed. More specifically we propose to build a set of indicators, inspired by the literature on diversity and coherence, but applicable by researchers using the proximity dimensions. It will offer indicators that are descriptive of the system individuals are participating in; the categories they are associated to (described by the 5 proximity dimensions); as well as the bridges/flows through a dynamic indicator of coherence, that capture the relationship occurring between individuals taking part in the research and the categories that they are associated to.

The contribution of the paper lies in the application of quantitative approaches to assess the diversity and coherence not only for cognitive aspects, but also to other social and organisational aspects in order to represent a wider range of aspects of research collaborations. The paper will first discuss how measures of diversity and coherence can be operationalised for application to the study of collaborative research using the proximities dimensions, and then proposes an operationalisation for each of the proximity.

THE GENERAL MEASURES

The measures of diversity and coherence are based upon the classification of *elements* into ***categories*** as part of a specific ***system*** (Rafols, 2014). In this paper we are using ***individuals*** as ***elements*** of a formal research collaboration (i.e. a team funded to undertake collaborative project supported by a research grant) which forms the overall ***system*** being studied. For each of the five proximities, individuals (elements) will be assigned to different positions (categories), and the distance between individuals is characterised by the gap or difference between categories. **Figure 1** shows a representation of both diversity and coherence (Source: Rafols 2014).

Figure 1: Illustration of the definitions of diversity (left) and



For example, consider two collaborators working together on a common paper, one from the University of Sussex and the other from the Universitat Politècnica de València. In geographical terms they will be assigned to different categories, one to Brighton and the other in València (our two positions/ categories defined by the towns they work in) and the distance between these positions can be defined by the travelling time required to meet each other (time is chosen in preference to distance, which does not take into account the transportation infrastructures that support collaboration). The distance is approximately 6 hours by combination of air and public transport.

The diversity measure describes three inter-related properties: the **variety** or number of categories, the **balance** of elements across categories and the **disparity** among categories. For example, following the geographic example above, a collaboration between two authors in the same city with another in a second city, is characterised as having less variety and balance than three authors in three cities, while the latter case has more disparity than the former.

Rao-Stirling diversity	$\sum_{i,j(i\neq j)} p_i p_j d_{ij}$

In addition to the positions of the elements, another property we are interested in is the relationship between actors. In the case of interactions, we need to take into account both the number of interactions and their intensity. This is captured by the coherence measure. The coherence measure describes three interrelated properties: the **density** of interactions, the relative **intensity** of the relations and the **disparity** across categories bridged by relations.

Coherence (γ=1, δ=1)	$\sum_{i,j(i\neq j)} i_{ij} d_{ij}$

Thus these two measures used along the five dimensions in the proximity framework capture the **position** of each element/individual within the system studied, but also the **bridges/flows** that emerged between the different dimensions. Later in the paper we will look at how we can apply these measures to study an illustrative case study of collaboration within a biomedical research project. For each indicator we can calculate position and flows before and after the research project started in order to capture the impact of the project on both the positions of the actors and the flows between actors. The paper follows on to describe further the operationalisation of the proximities in 1) the position and 2) the flows indicators.

The attribution of a specific position to an element (or individual) is made for each of the proximity. The indicator shows the distribution of the elements under study within specific sets of categories. The indicator is static as it provides a picture at a given time of the distribution of individuals across categories. To display the position of actors within categories we use an indicator based on the Rao-Stirling measure of diversity, which includes a component about the distance between individuals and the proportion of individuals within the categories in the system.

There are five categories, one for each dimension of Boschma's proximity framework: The social, the cognitive, the organisational, the institutional and the geographical. The five categories are here operationalised in five distinct attributes of distance. These require distinct strategies for operationalisation.

For three proximities (geographical, organisational, and institutional) the association of elements to each category is defined depending on where the individuals (elements) work (i.e. the geographical location of the workplace, the organisation employing the person and the type of institution) (Ponds et al. 2007, Hardeman et al. 2015). Another category is defined with reference to the knowledge base on which individuals build within the collaboration (the cognitive dimension) and is captured through the citations from scientific publication for each individual (Porter et al. 2007, Rafols 2014). The final proximity (the social dimension), assumes that each individual is both a distinct element and a different category. While we have described how the proposed operationalisation assigns elements to categories for the analysis, the diversity indicator also includes a distance component to represent disparity between each categories.

3. OPERATIONALISATION OF MEASURES OF DIVERSITY AND COHERENCE

In order to operationalise the measures, we need to compute the proportion of elements into categories (p_i) and define the proximities or distances between categories. This has to be done for each of the five analytical dimensions.

Social distance

For the ***social distance***, it is proposed to take into account whether individuals know each other before the project started. Two individuals will be considered close at the start of the project depending on two factors - whether they knew each other before the project (they would be considered quite close) and whether they worked together previously (in which case they will be considered very close - if they had worked together before and were prepared to work together again in a new project). We describe the distance between locations in terms of a categorical variable:

- Do not know (4/4)
- Know a bit (3/4)
- Have worked together before (0/4)

Cognitive distance

The ***cognitive distance*** is based on the journals cited in papers individuals considered authored. Each individual is associated with a number of journals they cite, and the distance between two individuals will be based upon the journals they cite and whether these journals are considered similar or not. The similarity between journals is defined by a similarity matrix, based on citation patterns for individual journals. This similarity matrix is produced on the basis of citations between the web of science indexed journal (for the last 5 year period), and has been kindly provided to us by the OST (Paris). This data is used to calculate cosines similarity between each pair of journals which is used as the similarity matrix between journals.

This journal similarity matrix is used together with citations patterns of individuals to calculate distances between individuals taking part in the project. This is performed using a method proposed by Zhou et al. (2012) who describe a way to compute a similarity-weighted cosine measure. The similarity-weighted cosine measure (which is in a normalised form) is defined as follow:

$$\varphi(\mathrm{X},\mathrm{Y}) = \frac{\varphi_{\mathrm{X},\mathrm{Y}}}{\sqrt{\varphi_{\mathrm{X}}\varphi_{\mathrm{Y}}}} = \frac{\sum_{i,j=1}^{N} p_{\mathrm{X}_i} p_{\mathrm{Y}_j} S_{ij}}{\sqrt{\left(\sum_{i,j=1}^{N} p_{\mathrm{X}_i} p_{\mathrm{X}_j} S_{ij}\right) \cdot \left(\sum_{i,j=1}^{N} p_{\mathrm{Y}_i} p_{\mathrm{Y}_j} S_{ij}\right)}}$$

This measure enables us to provide a similarity measure between two individuals depending on their cognitive background, which cited journals as a proxy.

Geographic distance

For the ***geographical distance***, individuals are assigned to a geographic location. The geographic location is assigned depending on the time spent by a person at a specific location. In some cases, this can be different from their affiliation (based on data reported by individuals, for example at interview). Thus the geographical and organisational distances can be based on different data, while organisations' addresses are used to calculate geographical distances. As previously noted we use travelling time as a proxy of geographical distance between two individuals working at different given locations.

We describe the distance between locations in terms of a categorical variable:

- Same department (3 minutes) (0/5)
- Same university, same campus (up to 15 minutes' walk) (1/5)
- Same city/metropolis (up to 2 hours) (2/5)
- Same region/country (up to 4-5 hours by train) (3/5)
- Same continent (flight or long train needed) (4/5)
- Other continent (5/5)

Organisational distance

For the ***organisational distance***, individuals are assigned to an organisation, the organisation they work in. There are different levels of organisational integration which we take into account when defining the distance (whether the individuals work in the same department or centre, or if an individual has a visiting status in an organisation).

We describe the distance between the organisations in terms of a categorical variable:

- Same department or centre (0/2)
- Same organisation (1/2)
- Different organisation (2/2)

If the person has a ***visiting status*** in an organisation he/she will have a (-1/2) to correct for the status, as these individuals may be considered closer in organisational terms than people who are completely external to the organisation.

Institutional distance

Finally, for the ***institutional distance***, we use previous literature in order to define distances between given institutions. As our example focuses on a biomedical research projects we consider six different type of institutions which has been previously identified in the literature (Rotolo et al., 2015): those involved in higher education/ research (e.g. universities), hospitals, governmental organisations, non-governmental organisations, industry, and university hospitals. In this latter category we differentiate between individuals mainly working as clinicians in university hospitals (referred to as working in Hosp/Univ) and those mainly

working as researchers (referred to as working in Univ/Hosp) given the different requirements attached to these roles. In order to identify distances between institutions, we consider the overlap over the general missions between these institutions. We consider whether these institutions' main objective are oriented towards commercialisation, Care, Open science, Education and Policy (Llopis & D'Este, 2016). The following table whether each institution has one or more of the following mission, with a yes or no answer represented by a binary attribute.

Table 1: Overlap of missions between different institutions

	Res&Edu	Hosp	GO	NGO	Industry	Univ/Hosp	Hosp/Univ
Commerc.	0	0	0	0	1	0	0
Care	0	1	0	0	0	1	1
Open Sc	1	1	1	1	0	1	1
Education	1	0	0	0	0	1	1
Policy	0	0	1	1	0	0	0

We use each columns as a vector of binary attributes (see table 1). The above table can then be interpreted as a contingency table for binary attributes. Using the symmetric binary dissimilarity method (Han, Kamber, & Pei, 2012, pp. 70–71) we can compute the table below :

Table 2: Institutional distance defined between pairs of institutions (1)

	Res & Edu	Hosp	GO	NGO	Industry	Univ/Hosp	Hosp/Univ
Res & Edu	0	0.4	0.4	0.4	0.6	0.2	0.2
Hosp	0.4	0	0.4	0.4	0.6	0.2	0.2
GO	0.4	0.4	0	0	0.6	0.6	0.6
NGO	0.4	0.4	0	0	0.6	0.6	0.6
Industry	0.6	0.6	0.6	0.6	0	0.8	0.8
Univ/Hosp	0.2	0.2	0.6	0.6	0.8	0	0
Hosp/Univ	0.2	0.2	0.6	0.6	0.8	0	0

The distance ranges from 0 to 0.8 between pairs of organisation. There are a few concerns with this similarity matrix as we would like to consider each institutions to be different from one another, thus GO and NGO must be superior to 0, and people working on the research side (Univ/Hosp) from the university-hospitals must be differentiated from those working as clinicians (Hosp/Univ). Also we would like to readjust the measures between Universities, Hosp, Univ/Hosp and Hosp/univ. As the primary focus of University-Hospitals and universities is teaching and open science, they should be closer than the ones that have their main focus on care (Hospitals and clinicians at university hospitals). Univ/Hosp and Hosp/Univ are different because the first is slightly more focused on open science and the latter is primarily focused on care. Thus the distance measure will be slightly modified to take into account this aspect.

FLOWS/BRIDGES (USING THE COHERENCE INDICATOR)

In addition to exploring ways to assess diversity, we also provide an operationalisation for assessing coherence that looks at the flows occurring within the project between the different categories across each of the five proximities. The bridges or flows are being represented

through the coherence indicator that include both factors for **distance** (which uses the distance measures discussed in the previous section) and **intensity** as introduced below.

The intensity of the flows are based on indicators of personal interactions made by individuals**.** The intensity measure is therefore defined by the frequency of interactions (i.e. whether these are daily, weekly, monthly, bi-annually, or yearly interactions). The measure of intensity is different to the social proximity set out above because here we are concerned not with how acquainted individuals are but by the frequency of interaction which is used as a proxy for intensity of collaboration.

Intensity measures

The scale of intensity can be derived from the frequency of the interaction, this is a measure of personal interaction . For example:

- no meeting (0)
- yearly meeting (1/5)
- every 6 month meeting (2/5)
- monthly meeting (3/5)
- weekly meeting (4/5)
- daily meeting (5/5)

4. APPLICATION TO A RESEARCH PROJECT ON PODOCONIOSIS

The indicators presented above were purposefully presented in a general manner in order to introduce a novel way to study collaboration. The last part of the paper aims at applying the developed indicators to a specific case. The case follows collaboration within a funded research project, in this case we will focus on a research project aiming at developing the understanding of a specific neglected disease, podoconiosis. Podoconiosis is a relatively under studied non-infectious neglected tropical disease which is characterised by the swelling of feet or lower part of the leg in affected individuals (Deribe, Tomczyk, & Tekola-Ayele, 2013). It is associated with social stigma and is also causes significant problems by reducing the economic activity of sufferers. The focal research project resulted in a substantial boost to the number of publications on this topic as well as increasing substantially the number of researchers working in this field.

The data relies on both publication and interview data gathered among individuals participating in the research project. Publication data were retrieved through the Web of Science and are mainly used to generate indicators and maps of cognitive proximity, as already explored in previous literature (Rafols, 2014). Interview data consists of gathering data about the other proximities such as organisations, institutions, geographical location (which can be crossed checked with the publication data), social relationships, but also data about intensity of interactions.

VISUALISATION

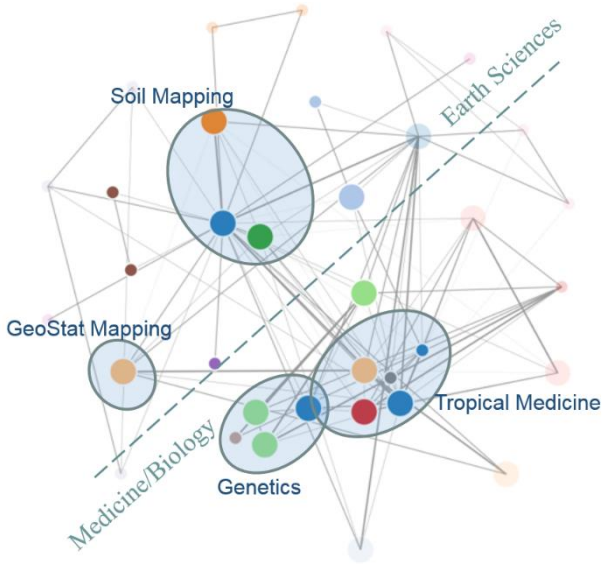
Diversity and coherence include different elements, such as distance between categories (included in both measures), the intensity of links for the coherence measures and the proportion of elements in each category (included in the diversity measure). For each of the measures, namely diversity and coherence, two of these aspects are integrated into the measure and therefore the analyst loose part of the complexity of the information held into the single metric. Thus the analysis of the proximity dimensions can be performed using both the measures (i.e. coherence and diversity) explained above, together with the visualisation that enables the user

to have a better understanding of the single metric. For instance, as discussed above, the diversity is based upon how elements are distributed into categories, and have attributes such as variety, balance, and disparity. In the same way coherence has properties such as density of interactions, intensity of relationships, and disparity across categories as well as how they are bridged.

The visualisations are represented in a two dimensional space and show distances between individual elements and links (intensity is displayed by the thickness of the line) between these individuals. The distribution of elements enables the analyst to identify categories. This can be done by using both the information given in interviews (for social and organisational proximities) and information held in raw data (for the cognitive side). Figure 2 to 7 shows such representations based on data collected in the Podoconiosis project case study, for each of the proximity dimensions using part of the metrics introduced above.

Cognitive proximity

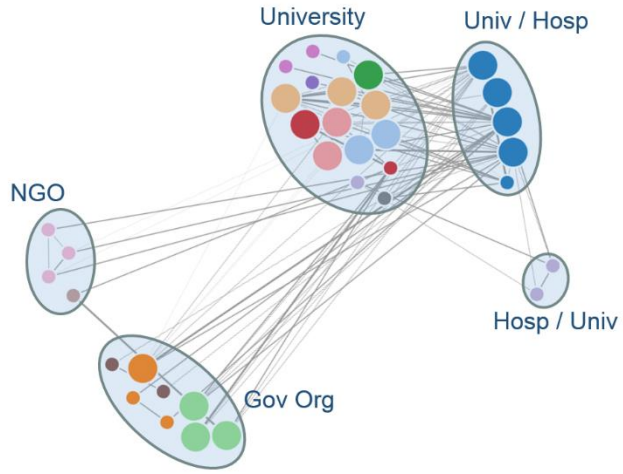
Figure 2: Links between individuals with cognitive node positioning



Diversity: 0.551
Coherence Before: 8.9
Coherence After: 21.65

Institutional proximity

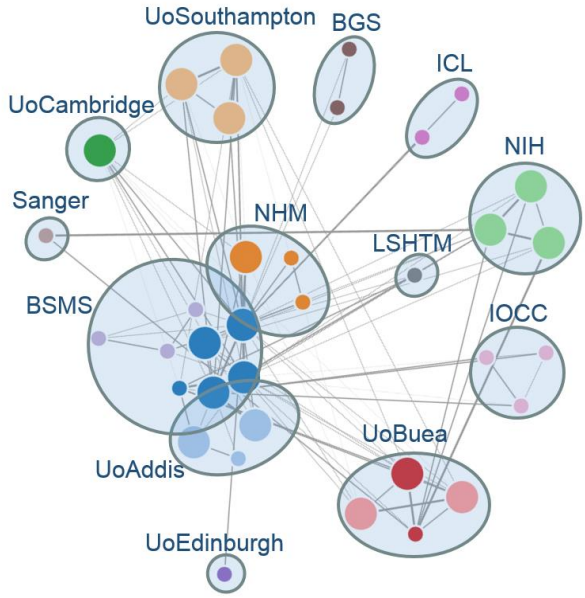
Figure 3: Links between individuals with institutional node positioning



Diversity: 0.31
Coherence Before: 7.44
Coherence After: 27.872

Organisational proximity

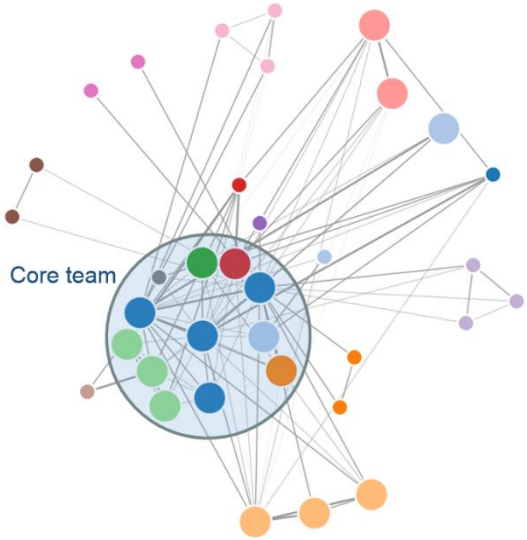
Figure 4: Links between individuals with organisational node positioning



Diversity: 0.905
Coherence Before: 20.6
Coherence After: 71

Social proximity

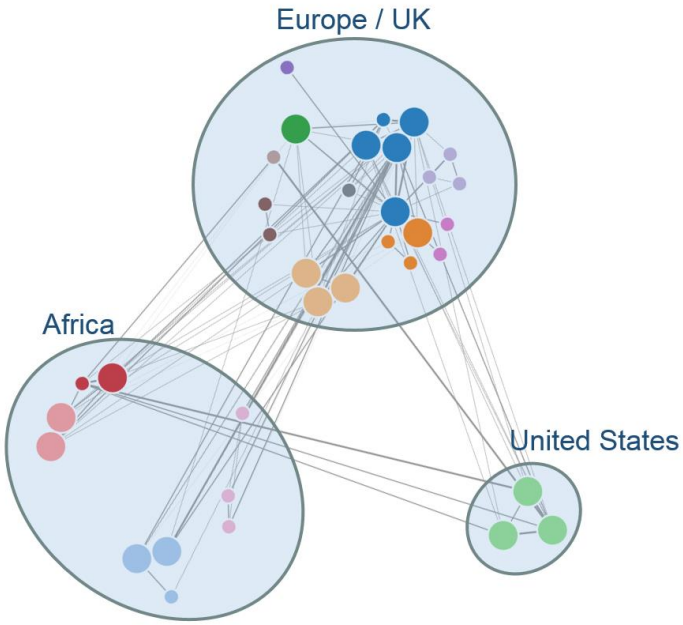
Figure 5: Links between individuals with node positioning



Diversity: 0.924
Coherence Before: 14
Coherence After: 77.1

Geographic proximity

Figure 6: Links between individuals with geographic node positioning



Diversity: 0.743
Coherence Before: 15.52
Coherence After: 57.12

5. CONCLUSIONS

The paper concludes with a discussion about the suitability of the proposed tool to assess potential knowledge integration through collaboration, and its strengths and limitations.

REFERENCES

Bordons, M., Morillo, F., & Gomez, I. (2004). Analysis of cross-disciplinary research through bibliometric tools. In: Moed, Glanzel, & Schmoch (Eds.), Handbook of quantitative science and technology research (pp. 437–456). Dordrecht: Kluwer.

Boschma, R. (2005). Proximity and Innovation: A Critical Assessment. *Regional Studies*, *39*(1), 61–74. doi:10.1080/0034340052000320887

Grant, R. M., & Baden-Fuller, C. (1995). A KNOWLEDGE-BASED THEORY OF INTER-FIRM COLLABORATION. *Academy of Management Proceedings*, *1995*(1), 17–21. doi:10.5465/AMBPP.1995.17536229

Han, J., Kamber, M., & Pei, J. (Computer scientist). (2012). *Data mining : concepts and techniques*. Elsevier/Morgan Kaufmann.

Llopis, O., & D'Este, P. (2016). Beneficiary contact and innovation: The relation between contact with patients and medical innovation under different institutional logics. *Research Policy*, *45*(8), 1512–1523. doi:10.1016/j.respol.2016.03.004

Maskell, P., & Malmberg, A. (1999). Localised learning and industrial competitiveness. *Cambridge Journal of Economics*, *23*(2), 167–185. doi:10.1093/cje/23.2.167

Nooteboom, B. (2000). Learning and innovation in organizations and economies.

Rafols, I. (2014). Measuring SKnowledge Integration and Diffusion: Measures and Mapping of Diversity and Coherencecholarly Impact. In Y. Ding, R. Rousseau, & D. Wolfram (Eds.), *Measuring Scholarly Impact* (pp. 169–190). Cham: Springer International Publishing. doi:10.1007/978-3-319-10377-8

Rafols, I., & Meyer, M. (2010). Diversity and network coherence as indicators of interdisciplinarity: Case studies in bionanoscience. *Scientometrics*, *82*(2), 263–287. doi:10.1007/s11192-009-0041-y

Rafols, I., Porter, A. L., & Leydesdorff, L. (2010). Science overlay maps: A new tool for research policy and library management. *Journal of the American Society for Information Science and Technology*, *61*(9), 1871–1887. doi:10.1002/asi.21368